PATENT ABSTRACTS OF JAPAN

(11)Publication number : 07-114572
(43)Date of publication of application : 02.05.1995

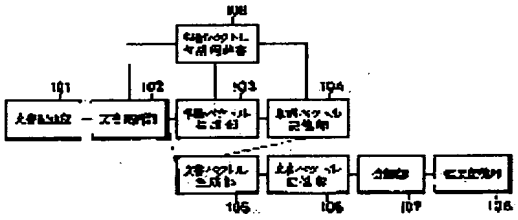
(51)Int.Cl. G06F 17/30
G06F 17/27

(21)Application number : 05-259809
(22)Date of filing : 18.10.1993

(71)Applicant : SHARP CORP
(72)Inventor : YUASA NATSUKI
UEDA TORU

(54) DOCUMENT CLASSIFYING DEVICE

(57)Abstract:
PURPOSE: To use semantic differences to automatically classify a document by automatically extracting feature vectors from the document and classifying the document based on these feature vectors.
CONSTITUTION: A storage part 101 where document data is stored, a document analysis part 102 which analyzes document data, a word vector generating part 103 which uses concurrent relations between words in the document to automatically generate a feature vector expressing the features of each word, a word vector storage part 104 where feature vectors are stored, a document vector generating part 105 which generates feature vectors of the document from feature vectors of words included in the document, a document vector storage part 106 where feature vectors of the document are stored, a classifying part 107 which uses the similarity between feature vectors of the document to classify the document, a result storage part 108 where the classification result is stored, and a feature vector generating dictionary 109 where words to be used for feature vector generation are registered are provided.



LEGAL STATUS

[Date of request for examination] 04.07.1997
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number] 2978044
[Date of registration] 10.09.1999
[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平7－114572

(43)公開日　平成７年(1995)５月２日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30					
17/27					
		9194－5L	Ｇ０６Ｆ　15/ 401	３１０　Ｄ	
		7315－5L	15/ 20	５５０　Ｆ	

審査請求　未請求　請求項の数３　ＯＬ　（全 10 頁）

(21)出願番号　特願平5－259809

(22)出願日　平成５年(1993)10月18日

(71)出願人　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町22番22号

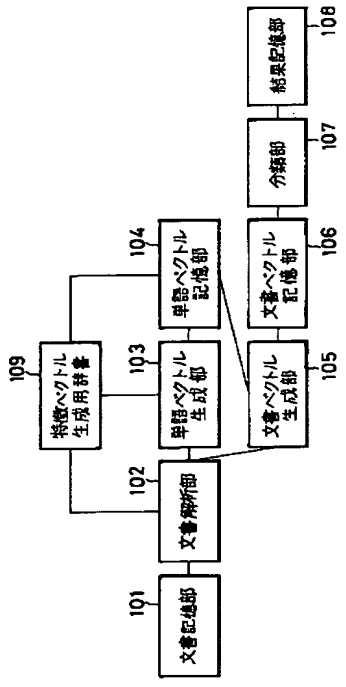
(72)発明者　湯浅　夏樹
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内

(72)発明者　上田　徹
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内

(74)代理人　弁理士　川口　義雄　（外１名）

(54)【発明の名称】　文書分類装置

(57)【要約】
【目的】　文書から自動的に単語の特徴ベクトルを抽出し、その特徴ベクトルをもとに文書を分類することで、意味的な異なりを用いた自動分類を可能にする。
【構成】　文書分類装置において、文書データを記憶する記憶部１０１と、文書データを解析する文書解析部１０２と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部１０３と、その特徴ベクトルを記憶する単語ベクトル記憶部１０４と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部１０５と、その特徴ベクトルを記憶する文書ベクトル記憶部１０６と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部１０７と、その分類した結果を記憶する結果記憶部１０８と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書１０９を備える。



1

【特許請求の範囲】

【請求項１】　文書分類装置において、文書データを記憶する記憶部と、文書データを解析する文書解析部と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部と、その特徴ベクトルを記憶する単語ベクトル記憶部と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部と、その特徴ベクトルを記憶する文書ベクトル記憶部と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部と、その分類した結果を記憶する結果記憶部と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書とを備え、大量の文書ファイル中の単語間の共起関係を用いて、各単語の特徴を表現する特徴ベクトルを自動的に生成し、文書を自動的に分類することができることを特徴とする文書分類装置。

【請求項２】　請求項１の文書分類装置の構成に加えて、結果記憶部に記憶されている分類結果を利用して分類時に有用な単語を選出する有用単語選出部を備え、大量の文書ファイルを分類した後でその分類された各分類群ごとに単語の出現率を調べることで、分類に有用な単語を選出し、分類に有用な単語のみを用いることで分類の精度を向上させることができることを特徴とする文書分類装置。

【請求項３】　請求項１あるいは請求項２の文書分類装置の構成に加えて、結果記憶部に記憶されている分類結果を利用して各分類群を代表する特徴ベクトルを求める代表ベクトル生成部と、代表ベクトル生成部で生成された代表ベクトルを記憶する代表ベクトル記憶部とを備え、大量の文書ファイルを分類した後でその分類された各分類群ごとの単語や文書の特徴ベクトルを用いて、その分野を代表する特徴ベクトルを求めることができることを特徴とする文書分類装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、文書を保存／自動分類する文書自動分類機やワープロ／ファイリングシステムなどに利用される文書分類装置に関する。

【０００２】

【従来の技術】従来、文書の自動分類は困難であり、ユーザが手動で分類を行なったり、文書中のキーワードを抽出し、あらかじめ作成されたシソーラスを用いて分類を行なっていた。また自動分類と称しているシステムでも分類のための基本的なデータは基本例文などの形で人手によって入力しておく必要があった。

【０００３】

【発明が解決しようとする課題】しかしながら、このような分類では人手による作業がボトルネックなるため、大量の文書の分類作業は大変困難である。

【０００４】本発明は以上の事情を考慮してなされたもので、人手を介することなく自動的に文書の分類を行なう文書分類装置を提供することを目的とする。

2

【０００５】

【課題を解決するための手段】請求項１に係る発明は、文書分類装置において、文書データを記憶する記憶部と、文書データを解析する文書解析部と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部と、その特徴ベクトルを記憶する単語ベクトル記憶部と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部と、その特徴ベクトルを記憶する文書ベクトル記憶部と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部と、その分類した結果を記憶する結果記憶部と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書とを備え、大量の文書ファイル中の単語間の共起関係を用いて、各単語の特徴を表現する特徴ベクトルを自動的に生成し、文書を自動的に分類することができることを特徴とする。

【０００６】また、請求項２に係る発明は、上記構成に加え、結果記憶部に記憶されている分類結果を利用して分類時に有用な単語を選出する有用単語選出部を更に備え、大量の文書ファイルを分類した後でその分類された各分類群ごとに単語の出現率を調べることで、分類に有用な単語を選出し、分類に有用な単語のみを用いることで分類の精度を向上させることができることを特徴とする。

【０００７】また、請求項３に係る発明は、上記構成に加え、結果記憶部に記憶されている分類結果を利用して各分類群を代表する特徴ベクトルを求める代表ベクトル生成部と、代表ベクトル生成部で生成された代表ベクトルを記憶する代表ベクトル記憶部を更に備え、大量の文書ファイルを分類した後でその分類された各分類群ごとの単語や文書の特徴ベクトルを用いて、その分野を代表する特徴ベクトルを求めることができることを特徴とする。

【０００８】

【作用】請求項１での単語の特徴ベクトルの学習時の作用を説明する。文書記憶部に記憶されている大量の文書ファイルの内容が文書解析部に渡されて文の解析（形態素解析など）が行なわれ、単語ベクトル生成部で単語の共起関係や出現頻度等を分析して各単語の特徴ベクトルを生成する。こうして生成された単語の特徴ベクトルは単語ベクトル記憶部に記憶される。このようにして単語の特徴ベクトルの学習が行なわれる。特徴ベクトルを生成する単語は特徴ベクトル生成用辞書に登録されている単語に制限することで特徴ベクトルの記憶空間が巨大になりすぎるのを防ぐ。

【０００９】請求項１での文書の分類時の作用を説明する。文章の分類を行なう時には、文書記憶部に記憶され

ている文書ファイルの内容が文書解析部に渡されて文の解析（形態素解析など）が行なわれ、文書ベクトル生成部では文書解析部で文の解析をした時に出現する単語の特徴ベクトルを単語ベクトル記憶部を参照して求め、文書に含まれる単語の特徴ベクトルから文書の特徴ベクトルを生成する。こうして生成された文書の特徴ベクトルは文書ベクトル記憶部に記憶され、この文書の特徴ベクトル間の類似度によって分類部で文書を分類する。この分類結果は結果記憶部に記憶される。

【００１０】請求項２に記載の構成では、大量の文書の分類を実行した後、有用単語選出部にて、結果記憶部に記憶されている分類結果を利用して分類時に有用な単語を選出する。有用単語選出部によって選出された単語だけを特徴ベクトル生成用辞書に登録してから再び単語の特徴ベクトルの学習を行なわせ、そうして得られた単語の特徴ベクトルを用いて再び分類を行なうことによって、請求項１の構成よりも特徴ベクトルの記憶空間を削減したり、また分類の精度をあげることもできる。

【００１１】請求項３に記載の構成では、大量の文書の分類を実行した後、代表ベクトル生成部にて、結果記憶部に記憶されている分類結果を利用して各分類群を代表する特徴ベクトルを求める。代表ベクトル生成部で生成された代表ベクトルは代表ベクトル記憶部に記憶される。一度各分類群の代表ベクトルを生成してしまえば、新たな文書データを分類する時には、その文書の特徴ベクトルと各分類群の代表ベクトルとの比較を行なうだけでその文書がどの分類群に属すかを判定できる。

【００１２】

【実施例】以下、本発明の好適な実施例を図面に基づき詳述する。

【００１３】請求項１に係る発明の一実施例を図１に示す。文書分類装置は、文書データを記憶する記憶部１０１と、文書データを解析する文書解析部１０２と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部１０３と、その特徴ベクトルを記憶する単語ベクトル記憶部１０４と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部１０５と、その特徴ベクトルを記憶する文書ベクトル記憶部１０６と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部１０７と、その分類した結果を記憶する結果記憶部１０８と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書１０９とから構成される。

【００１４】一般に通常の文書に使用されている単語の数は非常に多いため、特徴ベクトルを作成する際に用いる単語を制限しておく方が現実的である。このために用いるのが特徴ベクトル生成用辞書１０９で、ここに登録されている単語のみを用いて単語の特徴ベクトルを作成することで、特徴ベクトルの記憶空間の巨大化を抑えることができる。

【００１５】図２は単語の特徴ベクトルの学習時のシステム構成を示し、単語の特徴ベクトルの学習時には、学習用の大量の文書データ文書記憶部１０１に記憶させておき、文書記憶部１０１から読み出した文書データは記事、段落、一文等の適当な塊ごとに文書解析部１０２に読み込まれ、文書解析部１０２でその文書データを解析して単語が抽出される。ここで抽出された単語列をもとにして単語ベクトル生成部１０３で単語の特徴ベクトルを生成し、単語ベクトル生成部１０３で生成された単語の特徴ベクトルは単語ベクトル記憶部１０４に記憶される。こうして単語の特徴ベクトルを学習する。

【００１６】図３は文書分類時のシステム構成を示し、文書の分類をする時には、分類する文書のデータを文書記憶部１０１に記憶させておき、文書記憶部１０１から読み出した文書データは分類を行なわせたい単位（例えば記事単位）ごとに文書解析部１０２に読み込まれ、文書解析部１０２でその文書データの解析をして単語が抽出される。ここで抽出された単語の特徴ベクトルを１０４の単語ベクトル記憶部の内容を参照して求める。通常は文書データの一つの単位（例えば一つの記事）から複数の単語が抽出されるがこの時には求められるすべての単語の特徴ベクトルの値を平均化することで文書の特徴ベクトルが計算される。この時、単純に平均化するのではなく、各特徴ベクトルをその出現頻度の逆数に応じて重み付けをしてから（例えば、大量の記事からその単語の出現している記事数を調査し、ｌｏｇ（全記事数／その単語が出現している記事数）をその単語の特徴ベクトルに掛けてから）平均化するとより良い値が得られる場合がある。

【００１７】文書の特徴ベクトルが求まったら従来のクラスタリングの手法を適用することで文書の分類を行なうことができる。これは例えば文書の特徴ベクトル間の距離が近い文書同士は同じ分野に属するとみなせば良い。

【００１８】また、人間が各分類群ごとに典型的な文書を選び、その文書から抽出される単語の特徴ベクトルからその分類群の仮の代表ベクトルを生成しておき、文書記憶部１０１から読み込まれる文書の特徴ベクトルがどの分類群の仮の代表ベクトルに近いかで文書を分類することもできる。このような分類手法でも文書記憶部１０１から大量に文書データを読み込ませれば仮の代表ベクトルを人間が選んでいるという誤差の影響が少なくなり、最終的には各分野毎のかなり一般的な代表ベクトルを生成することができる。

【００１９】では具体的に単語の特徴ベクトルの生成法を説明する。単語の特徴ベクトルは、一塊の文書データの中に含まれている単語の出現頻度分布に、その単語のその一塊の文書データ中での出現頻度を掛けたものを加算していくことによって得られる。具体的な例で説明す

る。

【００２０】例文Ａ「アメリカ政府が先進主要国にココム規制の抜本的な見直しを提案してきた。」
例文Ｂ「規制対象国が兵器の製造につながる工業製品の輸出を規制することを条件に、ココムの規制品目を大幅に削減する意向のようだ。」
という文書データからどのように単語の特徴ベクトルを作成するかを説明する。ここでは、文書データは「一文」という単位で読み込まれることとするが、これは一記事など他の単位でも構わない。

【００２１】また特徴ベクトルの次元数が２１次元（特徴ベクトル生成用辞書に登録されている単語数が２１）で各要素が「アメリカ、政府、先進、主要、国、ココム、規制、抜本的、見直し、提案、対象、兵器、製造、工業、製品、輸出、条件、品目、大幅、削減、意向」という単語に対応しているとする。

【００２２】このような条件のもとで、例文Ａが文書記憶部１０１から読み込まれると、文書解析部１０２が解析されて「アメリカ、政府、先進、主要、国、ココム、規制、抜本的、見直し、提案」が抽出される。この時単語ベクトル生成部１０３ではこれらの単語すべての特徴ベクトルのこれらの単語に対応する要素に１を加算する。すると、「アメリカ」「政府」等、例文Ａに出現する単語の特徴ベクトルには（１，１，１，１，１，１，１，１，１，１，０，０，０，０，０，０，０，０，０，０）を加算する。これを図解したものが図８である。

【００２３】次に例文Ｂが文書記憶部１０１から読み込まれると、文書解析部１０２で解析されて、「規制、対象、国、兵器、製造、工業、製品、輸出、規制、条件、ココム、規制、品目、大幅、削減、意向」が抽出される。

【００２４】これから得られる単語出現頻度分布は（０，０，０，０，１，１，３，０，０，０，１，１，１，１，１，１，１，１，１，１）である。「規制」は３回出現しているので、この単語出現頻度分布を３倍したベクトルである（０，０，０，０，３，３，９，０，０，０，３，３，３，３，３，３，３，３，３，３）を「規制」の特徴ベクトルに加算し、「対象」「国」等、例文Ｂに１回しか出現しない単語の特徴ベクトルには（０，０，０，０，１，１，３，０，０，０，１，１，１，１，１，１，１，１，１，１）を加算する。これを図解したものが図９である。

【００２５】なお、このように常に整数を加算する方法では文の長さによって加算するベクトルの大きさが変化してしまうので、加算するベクトルの絶対値を１に正規化したり、出現頻度分布のベクトルの絶対値を１に正規化してから出現数に比例した値を掛けた後に加算する方法なども考えられる。

【００２６】そして最終的に得られた特徴ベクトルは、絶対値を１に正規化しておく。

【００２７】こうして得られた単語の特徴ベクトルは単語ベクトル記憶部１０４に記憶され、文書の分類時に利用される。

【００２８】次に、文書分類時の文書の特徴ベクトル生成の処理を、具体例として以下の例文Ｃが読み込まれた時をあげて説明する。

【００２９】例文Ｃ「アメリカ政府は兵器の削減を提案した。」
例文Ｃが文書記憶部１０１から読み込まれると、文書解析部１０２で解析されて「アメリカ、政府、兵器、削減、提案」が抽出される。すると文書ベクトル生成部１０５では単語ベクトル記憶部１０４の内容を参照して「アメリカ」「政府」等、例文Ｃに出現する単語の特徴ベクトルを加算していき、例文Ｃの特徴ベクトルとして（３，３，３，３，５，５，９，３，３，３，２，２，２，２，２，２，２，２，２，２）を得る。これを図解したものが図１０である。図１０ではわかりやすさを優先するためにベクトルの正規化を行なっていないが、実際の処理では加算する前に各単語の特徴ベクトルの絶対値を１に正規化してから加算を行なう。得られた特徴ベクトルは文書ベクトル記憶群１０６に記憶される。

【００３０】次に、分類時に分類部１０７にて文書の特徴ベクトルをどのように利用するのかを説明する。簡単には、まず求まった文書の特徴ベクトルの絶対値を１に正規化してから、Ｋ－ｍｅａｎｓ法などの従来からある手法を用いて分類したり、分類群の（仮）代表ベクトルとの類似度（距離を求めたり内積を計算することによって得られる）で分類すれば良いが、本手法で得られる特徴ベクトルは「多く出現する単語に対応する要素の値が非常に大きくなる」という特徴があるため、この特徴が分類に悪影響を与えないように工夫した方が良い分類結果が得られる場合が多い。例えば距離を求める場合には要素間の差が拡大しないような計算による距離（通常は各要素間の差の自乗和の平方根を計算するが、例えば各要素間の差の絶対値の和を計算するなどして求めた距離）を利用するようにしたほうが良いし、内積を求める前に全要素をｌｏｇをとったり冪乗根をとったりしてから正規化して値を均してから計算すると良い。

【００３１】分類の具体例として、分類群が３つあり、それぞれの分類群の代表ベクトルが以下のように求められていたとしよう。

【００３２】分類群１の代表ベクトル（１，１，１，１，０，０，０，０，０，０，０，０，０，０，０，０，１，１，１，１）
分類群２の代表ベクトル（１，１，１，１，１，１，１，１，１，５，５，５，５，５，５，５，５，５，５，５）
分類群３の代表ベクトル（４，４，４，４，６，６，

6，3，3，3，1，1，1，1，1，1，1，1，1，1，1）
類似度の尺度として、文書の特徴ベクトル、分類群の代表ベクトル共に絶対値を１に正規化してから両者の内積を計算し、一番大きな値をとるものが一番類似度が高いとすると、例文Ｃの特徴ベクトル
【００３３】
【数１】

$$\frac{1}{\sqrt{238}}$$

【００３４】（3，3，3，3，5，5，9，3，3，3，2，2，2，2，2，2，2，2，2，2，2）
分類群１の代表ベクトル
【００３５】
【数２】

$$\frac{1}{\sqrt{8}}$$

【００３６】（1，1，1，1，0，0，0，0，0，0，0，0，0，0，0，0，1，1，1，1）
分類群２の代表ベクトル
【００３７】
【数３】

$$\frac{1}{\sqrt{285}}$$

【００３８】（1，1，1，1，1，1，1，1，1，1，5，5，5，5，5，5，5，5，5，5）
分類群３の代表ベクトル
【００３９】
【数４】

$$\frac{1}{\sqrt{210}}$$

【００４０】（4，4，4，4，6，6，6，3，3，3，1，1，1，1，1，1，1，1，1，1）
なので、例文Ｃの特徴ベクトルと各分類群の代表ベクトルとの内積は分類群１との内積
【００４１】
【数５】

$$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{8}} \cdot 20 \simeq 0.4583$$

【００４２】分類群２との内積
【００４３】
【数６】

$$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{285}} \cdot 150 \simeq 0.5759$$

【００４４】分類群３との内積
【００４５】
【数７】

$$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{210}} \cdot 211 \simeq 0.9438$$

【００４６】となり、例文Ｃの特徴ベクトルは分類群３の代表ベクトルに一番近いことがわかるので、例文Ｃは分類群３に分類される。これを図解したのが図１１である。図１１も図１０と同様わかりやすさを優先するためにベクトルの正規化を行なっていないが、実際の処理では比較を行なう前に各ベクトルの絶対値を１に正規化してから比較を行なう。分類した結果は結果記憶部１０８に記憶される。
【００４７】次に、本発明の請求項２の一実施例を図４に示す。ここで、符号２０１～２０９で表されるものは図１の符号１０１～１０９で表されるものと夫々同じものである。
【００４８】文書分類装置は、文書データを記憶する記憶部２０１と、文書データを解析する文書解析部２０２と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部２０３と、その特徴ベクトルを記憶する単語ベクトル記憶部２０４と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部２０５と、その特徴ベクトルを記憶する文書ベクトル記憶部２０６と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部２０７と、その分類した結果を記憶する結果記憶部２０８と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書２０９と、結果記憶部２０８に記憶されている分類結果を利用して分類時に有用な単語を選出する有用単語選出部２１０とからなる。
【００４９】図５は学習時及び分類時のシステム構成を示す図である。最初は請求項１の実施例と同様の方法によって、単語の特徴ベクトルを学習し、それをもとに大量の文書データを分類する。分類した結果は結果記憶部２０８に記憶されているが、この結果を元にして、有用単語選出部２１０で有用単語の選出を行なう。これは、分類群ごとに各単語の頻度を求め、どの分類群にも同じような割合で含まれている単語を除去したり（方法１：最高頻度と最低頻度との比がある閾値以下のものを除去）、ある分類群にだけ高い割合で含まているものを選出したり（方法２：最高頻度と第二位頻度との比がある閾値以上のものを選出）する。なお、有用単語選出部２１０で選出を行なう単語は必ずしも特徴ベクトル生成用辞書２０９に登録されている単語からでなくても良く、もっと広い範囲の単語から選出を行なうことができる。
【００５０】具体例として分類群がａ，ｂ，ｃの三つあったとして、特徴ベクトル生成用辞書２０９に登録されている単語が「政治、日本、国際」の三つだったとする。そして分類群ごとに各単語（特徴ベクトル生成用辞書２０９に登録されている単語以外に「選挙」「問題」についても頻度を調べるとする）の頻度が次のようだったとする。
【００５１】分類群ａ　政治３０％，日本　５％，国際３５％，選挙１０％，問題２０％

分類群ｂ　政治　３％，日本５５％，国際３５％，選挙　２％，問題　５％
分類群ｃ　政治　３％，日本３０％，国際３５％，選挙　２％，問題３０％
すると、方法１を用いると「国際」はどの分類群にも同じような割合で含まれているので、特徴ベクトル生成用辞書から除去することになる。「政治」「日本」「選挙」「問題」は分類群ごとの頻度に偏りがあるので、有用単語として選出され、特徴ベクトル生成用辞書２０９に登録する（この時登録単語数を抑えたい場合は、頻度に偏りのある単語の中で、合計の出現頻度の順番に登録したい個数だけ取ってくれば良い）。方法２を用いた場合「政治」と「選挙」だけが選出され特徴ベクトル生成用辞書２０９に登録し、「日本」や「国際」や「問題」は特徴ベクトル生成用辞書２０９には登録しない。方法１と方法２の中間的な方法として、第１位の頻度と第ｎ位（ｎは３以上、分類群の個数－１以下）の頻度との比がある閾値以上であるかどうかで有用単語を選出する方法も考えられる。また、頻度の比ではなく、頻度の分散の値が大きいものを選出する方法も考えられる。

【００５２】なお、このようにして選出された単語は頻度の比（あるいは頻度の分散）に応じた重要度を持っていると考えることができるので、文書の特徴ベクトルを計算する時にはその文書内の単語の特徴ベクトルをこの比（あるいは分散）に応じて重み付けをしてから（例えば、ｌｏｇ（頻度の比）をその特徴ベクトルに掛けてから）平均化するとより良い文書の特徴ベクトル地が得られる場合がある。

【００５３】こうして特徴ベクトル生成用辞書２０９に、分類に有用な単語だけを登録し、もう一度、単語の特徴ベクトルを学習し、それを用いて文書を分類すると、特徴ベクトル生成用辞書をより小さくできたり、分類の精度をあげることができる。

【００５４】本発明の請求項３の一実施例を図６に示す。ここで、符号３０１～３１０で表されるものは図４の２０１～２１０で表されるものと夫々同じものである。

【００５５】文書分類装置は、文書データを記憶する記憶部３０１と、文書データを解析する文書解析部３０２と、文書中の単語間の共起関係を用いて各単語の特徴を表現する特徴ベクトルを自動的に生成する単語ベクトル生成部３０３と、その特徴ベクトルを記憶する単語ベクトル記憶部３０４と、文書内に含まれている単語の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部３０５と、その特徴ベクトルを記憶する文書ベクトル記憶部３０６と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部３０７と、その分類した結果を記憶する結果記憶部３０８と、特徴ベクトル生成時に使用する単語が登録されている特徴ベクトル生成用辞書３０９と、結果記憶部３０８に記憶されている分類結果を利用して分類時に有用な単語を選出する有用単語選出部３１０と、結果記憶部３０８に記憶されている分類結果を利用して各分類群を代表する特徴ベクトルを求める代表ベクトル生成部３１１と、代表ベクトル生成部３１１で生成された代表ベクトルを記憶する代表ベクトル記憶部３１２とからなる。

【００５６】なお請求項１の実施例を用いて請求項３のシステムを構成する場合には有用単語選出部３１０が無いシステムとなる。

【００５７】図７は学習時及び分類時のシステム構成を示す図である。最初は請求項１の実施例や請求項２の実施例と同様の方法によって、単語の特徴ベクトルを学習し、それをもとに大量の文書データを分類する。分類した結果は結果記憶部３０８に記憶されているが、この結果を元にして、代表ベクトル生成部３１１で代表ベクトルを生成する。これは例えば、分類群ごとの各単語の頻度を求め、ある分類群にだけ高い割合で含まれている単語を選出し、このような単語の特徴ベクトルの平均をとることよって生成できる。具体例として分類群がａ，ｂ，ｃの三つあったとして、特徴ベクトル生成用辞書３０９に登録されている単語が「政治、国会、国際」の三つだったとする。そして分類群ごとの各単語の頻度が次のようだったとする。

【００５８】
分類群ａ　政治４０％，国会５０％，国際１０％
分類群ｂ　政治１０％，国会１０％，国際８０％
分類群ｃ　政治２０％，国会１０％，国際７０％
すると、分類群ａの代表ベクトルは、「政治」の特徴ベクトルと「国会」の特徴ベクトルの平均として与えられる。なお単なる平均ではなく、出現割合によって、重みをつけることも考えられる。例えば「政治」の出現頻度が「国会」の出現頻度の２倍なら、「政治」の特徴ベクトルの２倍と「国会」の特徴ベクトルとを加算し、３で割ったものを分類群ａの代表ベクトルとする等である。

【００５９】同様に分類群ａに分類された文書の特徴ベクトルの平均をとったものを分類群ａの代表ベクトルとする方法も考えられる。

【００６０】こうして、代表ベクトルが生成されたらそれを代表ベクトル記憶部３１２に記憶しておくことで、以後の文書の分類時にはこの代表ベクトルを参照することで、文書記憶部３０１から読み込まれた文書は、その文書の特徴ベクトルにもっとも類似した代表ベクトルに対応する分類群に分類することができるようになる。

【００６１】本発明は文書分類に用いるだけでなく、電子メールや電子ニュースを自動的に分類したり、電子メールの中や電子ニュースの中からユーザーの興味を持ちそうなものを選出したり（ユーザーがそれまでに読んだメールやニュースの特徴ベクトルとの類似度で判定できる）、あいまい検索（検索キーワードの特徴ベクトルと、検索対象文書の特徴ベクトルとの類似度が一定の閾

値以上になる文書を検索するようにすることで、検索キーワードに正確にマッチしていなくても関連のキーワードで検索できる）に利用できたり、仮名漢字変換における同音意義語の選択（それまでに変換した内容から得られる特徴ベクトルとの類似度で同音意義語を選択する）に利用できたり、音声認識・手書き文字認識などにおいて過去の文脈に最も適合した変換結果を選択する方法をとる（それまでに認識した内容から得られる特徴ベクトルとの類似度で認識結果を選択する）際にも利用できたり、認識時等において単語等の検索空間を狭める（それまでに認識した内容から得られる特徴ベクトルの要素のうち一定の閾値以上になっている要素に対応する単語だけを検索するようにする）際にも利用できる。

【００６２】

【効果】本発明により、自動的に単語の特徴ベクトルを作成することができ、文書の分類を自動的に行なうことができるようになる。またこの方法で作成されたれ単語の特徴ベクトルは文書の分類時だけでなく、あいまい検索や、仮名漢字変換における同音意義語の選択にも利用できるし、音声認識・手書文字認識などにおいて、過去の文脈に最も適合した認識結果を選択する方法をとる際にも利用できる。

【図面の簡単な説明】

【図１】請求項１に係る発明の一実施例の基本構成を示すブロック図である。

【図２】図１に示すシステムの学習時のシステム構成を示すブロック図である。

【図３】図１に示すシステムの分類時のシステム構成を示すブロック図である。

＊【図４】請求項２に係る発明の一実施例の基本構成を示すブロック図である。

【図５】図４に示すシステムの学習、分類時のシステム構成を示すブロック図である。

【図６】請求項３に係る発明の一実施例の基本構成を示すブロック図である。

【図７】図６に示すシステムの学習、分類時のシステム構成を示すブロック図である。

【図８】単語の特徴ベクトルの生成を説明する図である。

【図９】単語の特徴ベクトルの生成を説明する図である。

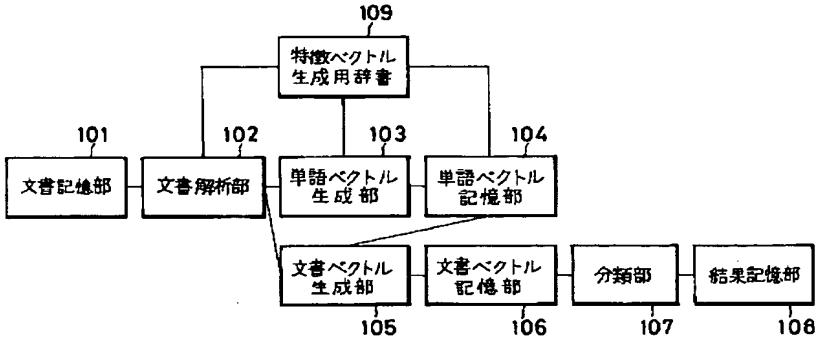
【図１０】文書の特徴ベクトルの生成を説明する図である。

【図１１】文書の分類を説明する図である。

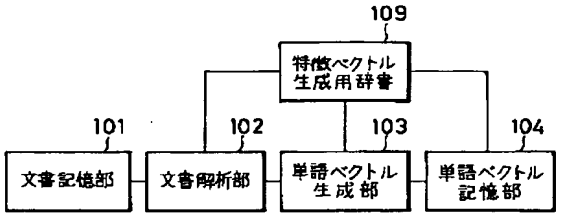
【符号の説明】

１０１、２０１、３０１　文書記憶部
１０２、２０２、３０２　文書解析部
１０３、２０３、３０３　単語ベクトル生成部
１０４、２０４、３０４　単語ベクトル記憶部
１０５、２０５、３０５　文書ベクトル生成部
１０６、２０６、３０６　文書ベクトル記憶部
１０７、２０７、３０８　分類部
１０８、２０８、３０８　結果記憶部
１０９、２０９、３０９　特徴ベクトル生成用辞書
２１０、３１０　有用単語選出部
３１１　代表ベクトル生成部
３１２　代表ベクトル記憶部

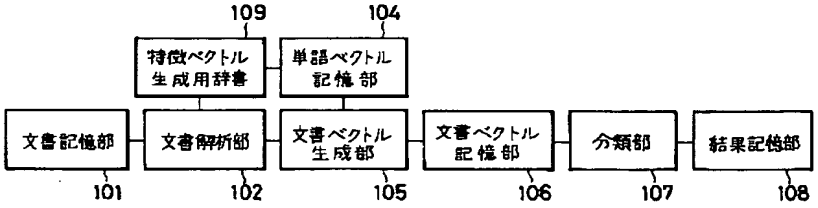
【図１】



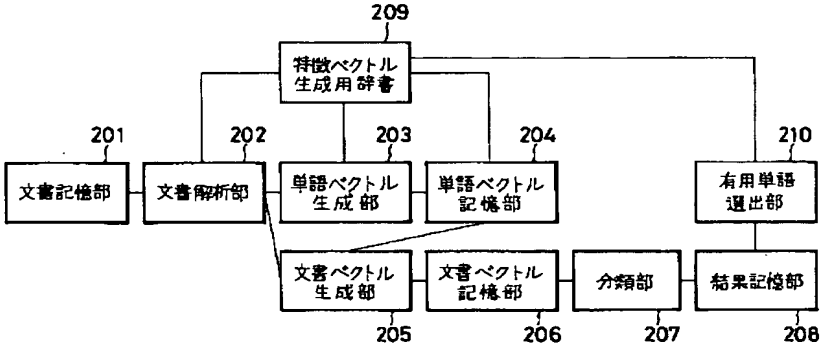
【図２】


109
特徴ベクトル生成用辞書
101
102
103
104
文書記憶部
文書解析部
単語ベクトル生成部
単語ベクトル記憶部

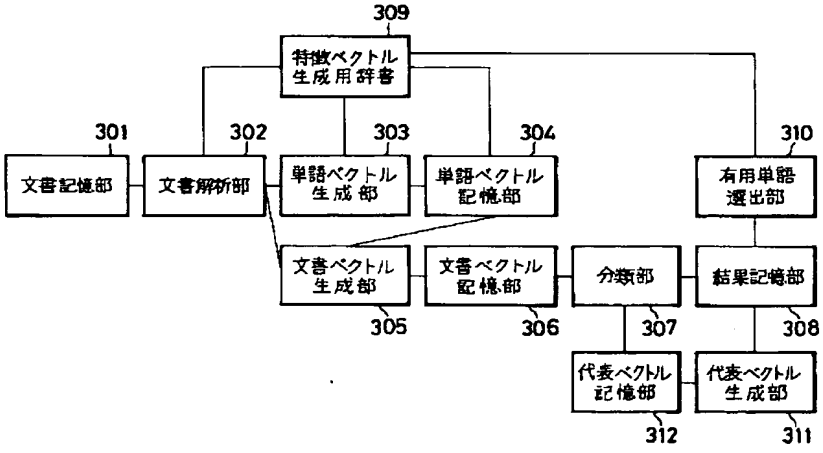
【図３】


109
104
特徴ベクトル生成用辞書
単語ベクトル記憶部
文書記憶部
文書解析部
文書ベクトル生成部
文書ベクトル記憶部
分類部
結果記憶部
101
102
105
106
107
108

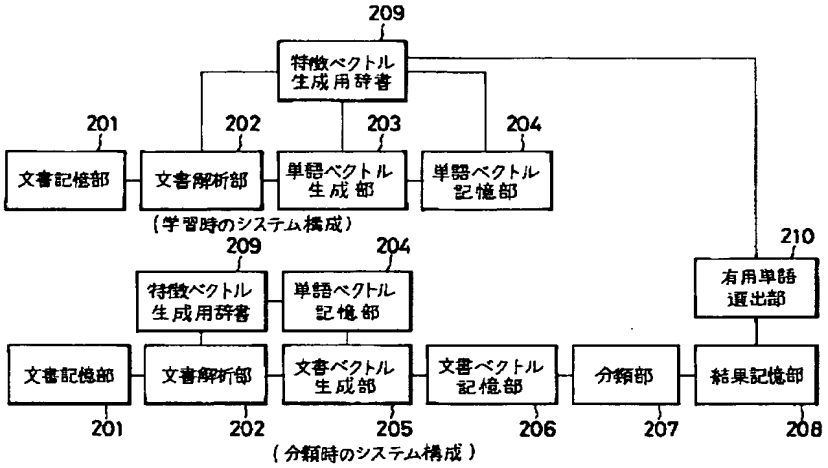
【図４】


209
特徴ベクトル生成用辞書
201
202
203
204
210
文書記憶部
文書解析部
単語ベクトル生成部
単語ベクトル記憶部
有用単語選出部
文書ベクトル生成部
文書ベクトル記憶部
分類部
結果記憶部
205
206
207
208

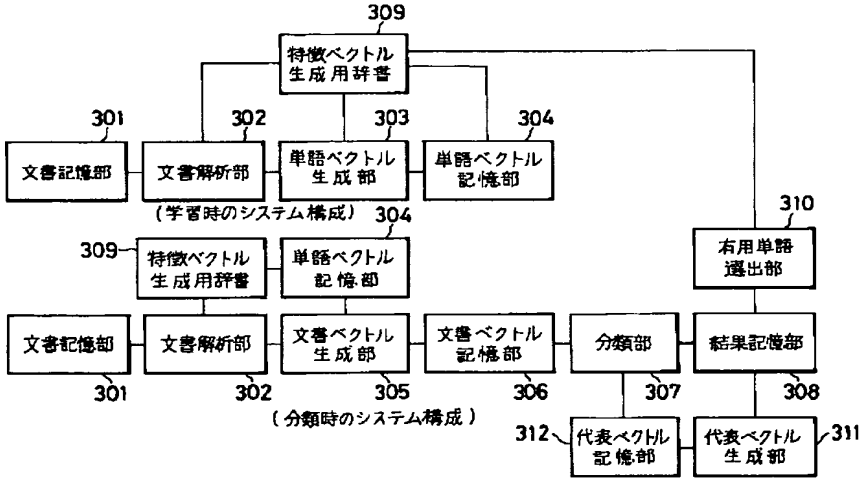
【図６】


309
特徴ベクトル生成用辞書
301
302
303
304
310
文書記憶部
文書解析部
単語ベクトル生成部
単語ベクトル記憶部
有用単語選出部
文書ベクトル生成部
文書ベクトル記憶部
分類部
結果記憶部
305
306
307
308
代表ベクトル記憶部
代表ベクトル生成部
312
311

【図５】



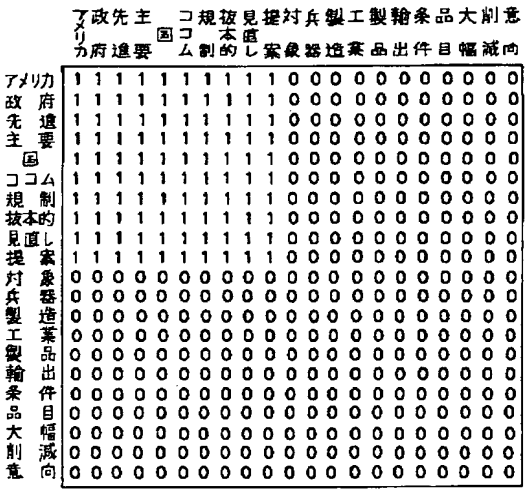
【図７】



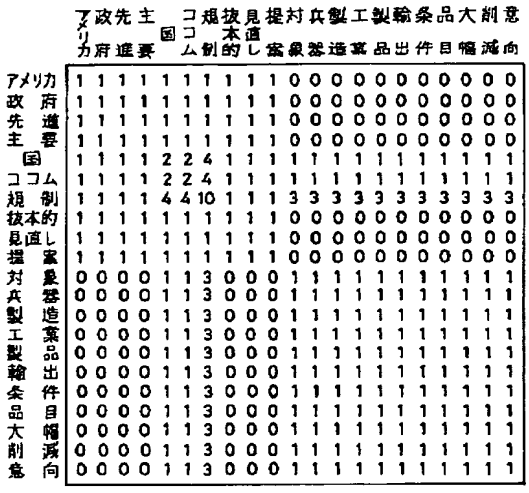
【図１０】

アメリカ	(	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0)
政　府	(	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0)
兵　器	(	0	0	0	0	1	1	3	0	0	0	1	1	1	1	1	1	1	1	1	1)
削　減	(	0	0	0	0	1	1	3	0	0	0	1	1	1	1	1	1	1	1	1	1)
+)　提　案	(	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0)
例文Cの特徴ベクトル	(	3	3	3	3	5	5	9	3	3	3	2	2	2	2	2	2	2	2	2	2)

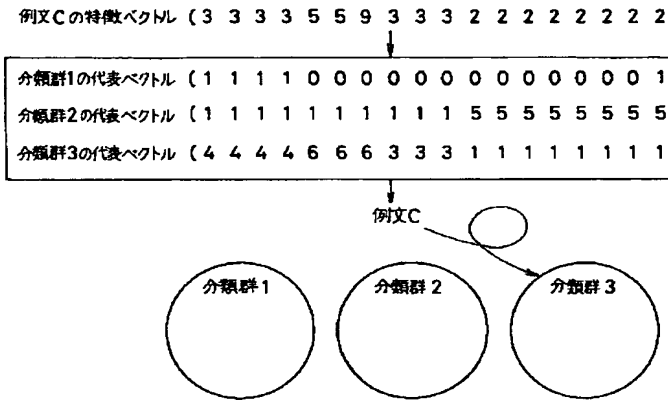
【図８】



【図９】



【図１１】


例文Cの特徴ベクトル（3 3 3 3 5 5 9 3 3 3 2 2 2 2 2 2 2 2 2 2 2）
分類群1の代表ベクトル（1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1）
分類群2の代表ベクトル（1 1 1 1 1 1 1 1 1 1 5 5 5 5 5 5 5 5 5 5 5）
分類群3の代表ベクトル（4 4 4 4 6 6 6 3 3 3 1 1 1 1 1 1 1 1 1 1 1）
例文C
分類群1
分類群2
分類群3

JAPANESE [JP,07-114572,A]

CLAIMS DETAILED DESCRIPTION TECHNICAL FIELD PRIOR ART EFFECT OF THE INVENTION TECHNICAL PROBLEM MEANS OPERATION EXAMPLE DESCRIPTION OF DRAWINGS DRAWINGS

[Translation done.]

CLAIMS

[Claim(s)]
[Claim 1] The storage section which memorizes document data in document classification equipment, and the document analysis section which analyzes document data, The word vector generation section which generates automatically the feature vector which expresses the description of each word using the coincidence relation between the words in a document, The document vector generation section which generates the feature vector of a document from the feature vector of the word vector storage section which memorizes the feature vector, and the word contained in the document, The document vector storage section which memorizes the feature vector, and the classification section which classifies a document using the similarity between the feature vectors of a document, As a result of memorizing the classified result, it has the storage section and the dictionary for feature-vector generation in which the word used for a feature-vector generate time is registered. Document classification equipment characterized by the ability to generate the feature vector expressing the description of each word automatically using the coincidence relation between the words in a lot of text files, and classify a document automatically.
[Claim 2] In addition to the configuration of the document classification equipment of claim 1, it has the useful word election section which elects a useful word using the classification result memorized by the result storage section at the time of a classification. Document classification equipment characterized by the ability to raise the precision of a classification by electing a useful word as a classification and using only a useful word for it at a classification by investigating the incidence of a word for each [which was classified] of that taxon of every after classifying a lot of text files.
[Claim 3] In the configuration of the document classification equipment of claim 1 or claim 2, in addition, the representation vector generation section which asks for the feature vector which represents each taxon using the classification result memorized by the result storage section, It has the representation vector storage section which memorizes the representation vector generated in the representation vector generation section. Document classification equipment characterized by the ability to ask for the feature vector which represents the field using the word for every taxon and feature vector of a document which were classified after classifying a lot of text files.

[Translation done.]

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Industrial Application] This invention relates to the document classification equipment used for the document automatic card counting sorter which is similar by preservation/automatic one, a word processor/filling system, etc. in a document.
[0002]
[Description of the Prior Art] Conventionally, the automatic classification of a document was difficult, and the user classified manually, or extracted the keyword in a document, and was classifying using the thesaurus created beforehand. Moreover, the fundamental data for a classification also with the system called automatic classification needed to be inputted by the help in forms, such as a basic example.
[0003]
[Problem(s) to be Solved by the Invention] however, the activity according to a help by such classification -- a bottleneck -- a sake -- a lot of documents -- classifying is very difficult.
[0004] This invention was made in consideration of the above situation, and aims at offering the document classification equipment which classifies a document automatically through a help.
[0005]
[Means for Solving the Problem] The storage section invention concerning claim 1 remembers document data to be in document classification equipment, The word vector generation section which generates automatically the feature vector which expresses the description of each word as the document analysis section which analyzes document data using the coincidence relation between the words in a document, The document vector generation section which generates the feature vector of a document from the feature vector of the word vector storage section which memorizes the feature vector, and the word contained in the document, The document vector storage section which memorizes the feature vector, and the classification section which classifies a document using the similarity between the feature vectors of a document, As a result of memorizing the classified result, it has the storage section and the dictionary for feature-vector generation in which the word used for a feature-vector generate time is registered. The feature vector expressing the description of each word is automatically generated using the coincidence relation between the words in a lot of text files, and it is characterized by the ability to classify a document automatically.
[0006] Moreover, in addition to the above-mentioned configuration, invention concerning claim 2 is further equipped with the useful word election section which elects a useful word using the classification result memorized by the result storage section at the time of a classification. It is characterized by the ability to raise the precision of a classification by electing a useful word as a classification and using only a useful word for it at a classification by investigating the incidence of a word for each [which was classified] of that taxon of every, after classifying a lot of text files.
[0007] Moreover, the representation vector generation section which asks for the feature vector which represents each taxon using the classification result invention concerning claim 3 is remembered to be by the result storage section in addition to the above-mentioned

configuration, It has further the representation vector storage section which memorizes the representation vector generated in the representation vector generation section. After classifying a lot of text files, it is characterized by the ability to ask for the feature vector representing the field using the word for every taxon and feature vector of a document which were classified.

[0008]

[Function] The operation at the time of study of the feature vector of the word in claim 1 is explained. The contents of a lot of text files memorized by the document storage section are passed to the document analysis section, analyses (morphological analysis etc.) of a sentence are performed, coincidence relation, the frequency of occurrence, etc. of a word are analyzed in the word vector generation section, and the feature vector of each word is generated. In this way, the feature vector of the generated word is memorized by the word vector storage section. Thus, study of the feature vector of a word is performed. The storage space of a feature vector prevents becoming huge too much with restricting the word which generates a feature vector to the word registered into the dictionary for feature-vector generation.

[0009] The operation at the time of a classification of the document in claim 1 is explained. The contents of the text file memorized by the document storage section when classifying a text are passed to the document analysis section, analyses (morphological analysis etc.) of a sentence are performed, in the document vector generation section, it asks for the feature vector of the word which appears when analyzing a sentence in the document analysis section with reference to the word vector storage section, and the feature vector of a document is generated from the feature vector of the word contained in a document. In this way, the feature vector of the generated document is memorized by the document vector storage section, and classifies a document according to the similarity between the feature vectors of this document in the classification section. This classification result is memorized by the result storage section.

[0010] With a configuration according to claim 2, after performing a classification of a lot of documents, a useful word is elected in the useful word election section using the classification result memorized by the result storage section at the time of a classification. By classifying again using the feature vector of the word which was made to learn the feature vector of a word again after registering into the dictionary for feature-vector generation only the word elected by the useful word election section, then was obtained, the storage spaces of a feature vector can be reduced rather than the configuration of claim 1, and the precision of a classification can also be raised.

[0011] With a configuration according to claim 3, it asks for the feature vector which represents each taxon with the representation vector generation section using the classification result memorized by the result storage section, after performing a classification of a lot of documents. The representation vector generated in the representation vector generation section is memorized by the representation vector storage section. Once it generates the representation vector of each taxon, when classifying new document data, it can judge to which taxon the document belongs only by performing the comparison with the feature vector of the document, and the representation vector of each taxon.

[0012]

[Example] Hereafter, the suitable example of this invention is explained in full detail based on a drawing.

[0013] One example of invention concerning claim 1 is shown in drawing 1 . The storage section 101 document classification equipment remembers document data to be, and the document analysis section 102 which analyzes document data, The word vector generation section 103 which generates automatically the feature vector which expresses the description of each word using the coincidence relation between the words in a document, The document vector generation section 105 which generates the feature vector of a document from the feature vector of the word vector storage section 104 which memorizes the feature vector, and the word contained in the document, The document vector storage section 106 which memorizes the feature vector, and the classification section 107 which classifies a document using the similarity between the feature vectors of a document, As a result of memorizing the classified result, it

consists of the storage section 108 and a dictionary 109 for feature-vector generation in which the word used for a feature-vector generate time is registered.
[0014] It is more realistic to restrict the word used in case a feature vector is created, since there are very many words currently generally used for the usual document. For this reason, it is the dictionary 109 for feature-vector generation, and creating the feature vector of a word only using the word registered here uses, and it can suppress growing gigantic of the storage space of a feature vector.
[0015] Drawing 2 shows the system configuration at the time of study of the feature vector of a word, at the time of study of the feature vector of a word, a lot of document data document storage section 101 for study is made to memorize, the document data read from the document storage section 101 are read into the document analysis section 102 for every suitable lumps, such as a report, a paragraph, and one etc. sentence, the document data is analyzed in the document analysis section 102, and a word is extracted. The feature vector of the word which generated the feature vector of a word in the word vector generation section 103 based on the word train extracted here, and was generated in the word vector generation section 103 is memorized by the word vector storage section 104. In this way, the feature vector of a word is learned.
[0016] When drawing 3 shows the system configuration at the time of a document classification and a document is classified, the document data which the document storage section 101 was made to memorize the data of the document to classify, and read them from the document storage section 101 are read into every [to make it classify into] unit (for example, report unit) by the document analysis section 102, the document data is analyzed in the document analysis section 102, and a word is extracted. It asks for the feature vector of the word extracted here with reference to the contents of the word vector storage section of 104. Usually, although two or more words are extracted from one unit (for example, one report) of document data, at this time, the feature vector of a document is calculated by equalizing the value of the feature vector of all the words called for. a better value may be acquired, if it does not equalize simply, but each feature vector is equalized at this time after carrying out weighting according to the inverse number of that frequency of occurrence, investigating the number of reports in which that word has appeared from a lot of reports and hanging log (the number of reports in which the total number of reports / its word has appeared) on the feature vector of that word for example, --
[0017] If the feature vector of a document can be found, a document can be classified according to applying the technique of the conventional clustering. What is necessary is just to consider that documents with a distance near [this] between the feature vectors of a document belong to the same field.
[0018] Moreover, human being chooses a typical document for every taxon, the temporary representation vector of the taxon is generated from the feature vector of the word extracted from the document, and the feature vector of the document read from the document storage section 101 can also classify a document according to whether to be close to the temporary representation vector of which taxon. If document data are made to read from the document storage section 101 in large quantities also by such classification technique, the effect of the error that human being has chosen the temporary representation vector decreases, and, finally the quite general representation vector for each field can be generated.
[0019] Then, the method of generating the feature vector of a word is explained concretely. The feature vector of a word is obtained by adding what applied the frequency of occurrence in the inside of the document data of the lump of the word to frequency-of-occurrence distribution of the word contained in the document data of a lump. A concrete example explains.
[0020] Example A "the American government has proposed radical reexamination of the COCOM regulation to the advanced major power."
Example B "to which it seems that the country for regulation is inclined to reduce COCOM's regulated items sharply on condition that export of the industrial product which leads to manufacture of arms is regulated"
It explains how the feature vector of a word is created from the document data to say. Here although [document data] read in the unit of "one sentence", other units, such as one report,

are sufficient as this.
[0021] moreover, the number of dimension of a feature vector -- 21 dimensions (the number of words registered into the dictionary for feature-vector generation is 21) -- each element -- "United States, the government, advanced, main, a country, COCOM, and regulation -- it improves and suppose that radical and the word a proposal, an object, arms, manufacture, industry, a product, export, conditions, items, large, reduction, and intention" are supported.
[0022] under such conditions, if Example A is read from the document storage section 101, the document analysis section 102 will analyze -- having -- "United States, the government, advanced, main, a country, and COCOM -- radical [regulation and] -- it improves and proposal" is extracted. At this time, 1 is added to the element corresponding to these words of the feature vectors of all these words in the word vector generation section 103. Then, the "United States", the "government", etc. add (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) to the feature vector of the word which appears in Example A. The thing illustrating this is drawing 8 .
[0023] Next, if Example B is read from the document storage section 101, it will be analyzed in the document analysis section 102, and "regulation, an object, a country, arms, manufacture, industry, a product, export, regulation, conditions, COCOM, regulation, items, large, reduction, and an intention" will be extracted.
[0024] The word frequency-of-occurrence distribution acquired from now on is (0, 0, 0, 0, 1, 1, 3, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1). It adds to the feature vector of "regulation". the vector which doubled this word frequency-of-occurrence distribution three since "regulation" had appeared 3 times -- it is (0, 0, 0, 0, 3, 3, 9, 0, 0, 0, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3) -- An "object", a "country", etc. add (0, 0, 0, 0, 1, 1, 3, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1) to the feature vector of the word which appears in Example B only once. The thing illustrating this is drawing 9 .
[0025] In addition, since the magnitude of a vector added with the die length of a sentence by the approach of always adding an integer in this way changes, how to add, after normalizing the absolute value of a vector to add to 1, or normalizing the absolute value of a vector of frequency-of-occurrence distribution to 1 and applying the value proportional to the number of appearances etc. is considered.
[0026] And the feature vector finally obtained normalizes the absolute value to 1.
[0027] In this way, the feature vector of the obtained word is memorized by the word vector storage section 104, and is used at the time of a classification of a document.
[0028] Next, the time of the following examples C being read considering processing of feature-vector generation of the document at the time of a document classification as an example is raised and explained.
[0029] Example C "the American government proposed reduction of arms."
If Example C is read from the document storage section 101, it will be analyzed in the document analysis section 102, and "the United States, the government, arms, reduction, and a proposal" will be extracted. Then, in the document vector generation section 105, with reference to the contents of the word vector storage section 104, the "United States", the "government", etc. add the feature vector of the word which appears in Example C, and get (3, 3, 3, 3, 5, 5, 9, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2) as a feature vector of Example C. The thing illustrating this is drawing 10 . In drawing 10 , since priority is given to intelligibility, the normalization of a vector is not performed, but in actual processing, after normalizing the absolute value of the feature vector of each word to 1 before adding, it adds. The obtained feature vector is memorized by the document vector storage group 106.
[0030] Next, it explains how the feature vector of a document is used in the classification section 107 at the time of a classification. After normalizing simply the absolute value of the feature vector of the document which was able to be found first to 1 Although what is necessary is to classify using a certain technique from the former, such as the K-means method, or just to classify according to similarity (obtained by finding distance or calculating an inner product) with the representation (temporary) vector of a taxon Since the feature vector obtained by this technique has the description "the value of the element corresponding to the word appearing [many] becomes very large", a classification result with it better [to devise so that this

description may not have a bad influence on a classification] is obtained in many cases. For example, distance by count which the difference between elements does not expand in finding distance (although the square root of the sum of squares of the difference between each element is usually calculated) For example, it is good to normalize, after taking log for all elements or taking a power root, before it is better to have used the distance which calculated and asked for the sum of the absolute value of the difference between each element and asking for an inner product, and to calculate, after leveling a value.
[0031] The juniper currently asked for those with three, and the representation vector of each taxon for the taxon as follows as an example of a classification.
[0032] The representation vector of a taxon 1 (1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1)
The representation vector of a taxon 2 (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5)
The representation vector of a taxon 3 (4, 4, 4, 4, 6, 6, 6, 3, 3, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
After the feature vector of a document and the representation vector of a taxon normalize an absolute value to 1, supposing what calculates both inner product and takes the biggest value is the highest as a scale of similarity as for similarity, it is the feature vector [0033] of Example C.
[Equation 1]
$\frac{1}{\sqrt{238}}$

[0034] (3,3,3,3,5,5,9,3,3,3,2,2,2,2,2,2,2,2,2,2)
The representation vector of a taxon 1 [0035]
[Equation 2]
$\frac{1}{\sqrt{8}}$

[0036] (1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1)
The representation vector of a taxon 2 [0037]
[Equation 3]
$\frac{1}{\sqrt{285}}$

[0038] (1,1,1,1,1,1,1,1,1,1,5,5,5,5,5,5,5,5,5,5)
The representation vector of a taxon 3 [0039]
[Equation 4]
$\frac{1}{\sqrt{210}}$

[0040] (4,4,4,4,6,6,6,3,3,3,1,1,1,1,1,1,1,1,1,1)
since -- the inner product of the feature vector of Example C, and the representation vector of each taxon -- an inner product [0041] with a taxon 1
[Equation 5]
$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{8}} \cdot 20 \simeq 0.4583$

[0042] An inner product with a taxon 2 [0043]
[Equation 6]
$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{285}} \cdot 150 \simeq 0.5759$

[0044] An inner product with a taxon 3 [0045]
[Equation 7]
$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{210}} \cdot 211 \simeq 0.9438$

[0046] Since it turns out that the feature vector of a next door and Example C is the closest to the representation vector of a taxon 3, Example C is classified into a taxon 3. Drawing 11 illustrated this. Since priority is given to intelligibility like [drawing 11] drawing 10 , the

normalization of a vector is not performed, but in actual processing, after normalizing each absolute value of a vector to 1 before comparing, it compares. The classified result is memorized by the result storage section 108.
[0047] Next, one example of claim 2 of this invention is shown in drawing 4 . Here, what is expressed with signs 201-209 is the same as what is expressed with the signs 101-109 of drawing 1 respectively.
[0048] The storage section 201 document classification equipment remembers document data to be, and the document analysis section 202 which analyzes document data, The word vector generation section 203 which generates automatically the feature vector which expresses the description of each word using the coincidence relation between the words in a document, The document vector generation section 205 which generates the feature vector of a document from the feature vector of the word vector storage section 204 which memorizes the feature vector, and the word contained in the document, The document vector storage section 206 which memorizes the feature vector, and the classification section 207 which classifies a document using the similarity between the feature vectors of a document, As a result of memorizing the classified result, it consists of the storage section 208, a dictionary 209 for feature-vector generation in which the word used for a feature-vector generate time is registered, and the useful word election section 210 which elects a useful word using the classification result memorized by the storage section 208 the result at the time of a classification.
[0049] Drawing 5 is drawing showing the system configuration at the time of study and a classification. At first, the feature vector of a word is learned and a lot of document data are classified according to the same approach as the example of claim 1 based on it. Although the classified result is memorized by the result storage section 208, it carries out based on this result, and a useful word is elected in the useful word election section 210. This asks for the frequency of each word for every taxon, removes the word contained at the same rate as every taxon, or elects a ********** thing for the following [a threshold with the ratio of the (approach 1:highest frequency and the minimum frequency] only as removal) and a certain taxon at a high rate (approach 2: elect the thing beyond the highest frequency and a threshold with the second place of a ratio with frequency). In addition, the word which elects in the useful word election section 210 may not necessarily be from the word registered into the dictionary 209 for feature-vector generation, and can perform election from the word of the larger range.
[0050] As an example, a taxon presupposes that the word which are a, b, and c and which is registered into the dictionary 209 for feature-vector generation was three "politics, Japan, and international" noting that there are three. And the frequency of each word (suppose that frequency is investigated also about "Election" and a "problem" in addition to the word registered into the dictionary 209 for feature-vector generation) presupposes that it was as follows for every taxon.
[0051] Taxon a 30% of politics, Japan 5%, 35% of international, 10% of Election, 20% taxon b of problems Politics 3%, 55% of Japan, 35% of international, Election 2%, problem 5% taxon c Politics 3%, 30% of Japan, 35% of international, Election If it carries out 30% of problems 2% Since "international" is contained at the same rate as every taxon if an approach 1 is used, it will remove from the dictionary for feature-vector generation. Since "politics", "Japan", "Election", and a "problem" have a bias in the frequency for every taxon, it is elected as a useful word and registers with the dictionary 209 for feature-vector generation (what is necessary is to take only the number to register in order of the total frequency of occurrence in the word which has a bias in frequency to stop the number of registered words at this time). When an approach 2 is used, "politics" and "Election" are elected and it registers with the dictionary 209 for feature-vector generation, and "Japan", international ["international"], and a "problem" are not registered into the dictionary 209 for feature-vector generation. The method of electing a useful word by whether it is beyond a threshold with the ratio of the frequency of the 1st place and the frequency of the n-th place (n is the number of 3 or more and a taxon - one or less) as the in-between approach of an approach 1 and an approach 2 is also considered. Moreover, the method of electing what has large not the ratio of frequency but value of distribution of frequency is also considered.

[0052] In addition, since it is possible that the word elected by doing in this way has the significance according to the ratio (or distribution of frequency) of frequency, if the feature vector of the word in that document is equalized after carrying out weighting according to this ratio (or distribution) when calculating the feature vector of a document (after hanging log (ratio of frequency) on that feature vector), the feature-vector ground of a better document may be obtained.
[0053] In this way, only a useful word is registered into a classification, once again, if the feature vector of a word is learned and a document is classified using it, the dictionary for feature-vector generation can be made smaller, or the precision of a classification can be raised to the dictionary 209 for feature-vector generation.
[0054] One example of claim 3 of this invention is shown in drawing 6 . Here, what is expressed with signs 301-310 is the same as what is expressed with 201-210 of drawing 4 respectively.
[0055] The storage section 301 document classification equipment remembers document data to be, and the document analysis section 302 which analyzes document data, The word vector generation section 303 which generates automatically the feature vector which expresses the description of each word using the coincidence relation between the words in a document, The document vector generation section 305 which generates the feature vector of a document from the feature vector of the word vector storage section 304 which memorizes the feature vector, and the word contained in the document, The document vector storage section 306 which memorizes the feature vector, and the classification section 307 which classifies a document using the similarity between the feature vectors of a document, The dictionary 309 for feature-vector generation in which the storage section 308 and the word used for a feature-vector generate time are registered as a result of memorizing the classified result, The useful word election section 310 which elects a useful word using the classification result memorized by the result storage section 308 at the time of a classification, It consists of the representation vector generation section 311 which asks for the feature vector which represents each taxon using the classification result memorized by the result storage section 308, and the representation vector storage section 312 which memorizes the representation vector generated in the representation vector generation section 311.
[0056] In addition, in constituting the system of claim 3 using the example of claim 1, it becomes a system without the useful word election section 310.
[0057] Drawing 7 is drawing showing the system configuration at the time of study and a classification. At first, the feature vector of a word is learned and a lot of document data are classified according to the same approach as the example of claim 1, or the example of claim 2 based on it. Although the result storage section 308 memorizes, the classified result is carried out based on this result, and generates a representation vector in the representation vector generation section 311. this asking for the frequency of each word for every taxon, electing the word contained only in a certain taxon at a high rate, and taking the average of the feature vector of such a word -- it is generable. As an example, a taxon presupposes that the word which are a, b, and c and which is registered into the dictionary 309 for feature-vector generation was three "politics, Parliament, and international" noting that there are three. And the frequency of each word for every taxon presupposes that it was as follows.
[0058]
Taxon a 40% of politics, 50% of Parliaments, 10% taxon b of international 10% of politics, 10% of Parliaments, 80% taxon c of international If it carries out 20% of politics, 10% of Parliaments, and 70% of international, the representation vector of Taxon a will be given as an average of a "political" feature vector and the feature vector of "Parliament." In addition, giving weight is also considered by not a mere average but the appearance rate. For example, if the "political" frequency of occurrence is twice the frequency of occurrence of "Parliament", it is making into the representation vector of Taxon a what added the twice of a "political" feature vector, and the feature vector of "Parliament", and was divided by 3 etc.
[0059] How to make what took the average of the feature vector of the document similarly classified into Taxon a the representation vector of Taxon a is also considered.
[0060] In this way, the document read from the document storage section 301 can be classified

now into the taxon corresponding to a representation vector most similar to the feature vector of that document according to referring to this representation vector at the time of a classification of future documents by memorizing it in the representation vector storage section 312, if a representation vector is generated.

[0061] This invention classifies an electronic mail and electronic news automatically, or it not only uses it for a document classification, but Elect what is likely to have a user's interest out of an electronic mail and electronic news, or (A user can judge by similarity with the feature vector of the mail read by then or news) Ambiguous retrieval (by searching the document which becomes beyond a threshold with the fixed similarity of the feature vector of a retrieval keyword, and the feature vector of the document for retrieval) Even if it does not match a retrieval keyword correctly, can use for the ability to refer to the keyword of relation, or Can use for selection (the homophenes is chosen by similarity with the feature vector obtained from the contents changed by then) of the homophenes in the conversion of kana into kanji, or Also in case the approach of choosing the conversion result of having suited the past context most in speech recognition, handwriting recognition, etc. is taken (a recognition result is chosen by similarity with the feature vector obtained from the contents recognized by then), can use, or It can use, also in case retrieval space, such as a word, is narrowed in the time of recognition etc. (only the word corresponding to the element which has become among the elements of the feature vector obtained from the contents recognized by then beyond the fixed threshold is searched).

[0062]

[Effect] The feature vector of a word can be created automatically and a document can be automatically classified now according to this invention. Moreover, it is created by this approach and hangs down, and it can use not only for the time of a classification of a document but for selection of ambiguous retrieval and the homophenes in the conversion of kana into kanji, and in speech recognition, hand written character recognition, etc., also in case the feature vector of a word takes the approach of choosing the recognition result of having suited the past context most, it can be used.

[Translation done.]

TECHNICAL FIELD

[Industrial Application] This invention relates to the document classification equipment used for the document automatic card counting sorter which is similar by preservation/automatic one, a word processor/filling system, etc. in a document.

[Translation done.]

PRIOR ART

[Description of the Prior Art] Conventionally, the automatic classification of a document was difficult, and the user classified manually, or extracted the keyword in a document, and was classifying using the thesaurus created beforehand. Moreover, the fundamental data for a classification also with the system called automatic classification needed to be inputted by the help in forms, such as a basic example.

[Translation done.]

EFFECT OF THE INVENTION

[Effect] The feature vector of a word can be created automatically and a document can be automatically classified now according to this invention. Moreover, it is created by this approach and hangs down, and it can use not only for the time of a classification of a document but for selection of ambiguous retrieval and the homophenes in the conversion of kana into kanji, and in speech recognition, hand written character recognition, etc., also in case the feature vector of a word takes the approach of choosing the recognition result of having suited the past context most, it can be used.

[Translation done.]

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention] however, the activity according to a help by such classification -- a bottleneck -- a sake -- a lot of documents -- classifying is very difficult.
[0004] This invention was made in consideration of the above situation, and aims at offering the document classification equipment which classifies a document automatically through a help.

[Translation done.]

MEANS

[Means for Solving the Problem] The storage section invention concerning claim 1 remembers document data to be in document classification equipment, The word vector generation section which generates automatically the feature vector which expresses the description of each word as the document analysis section which analyzes document data using the coincidence relation between the words in a document, The document vector generation section which generates the feature vector of a document from the feature vector of the word vector storage section which memorizes the feature vector, and the word contained in the document, The document vector storage section which memorizes the feature vector, and the classification section which classifies a document using the similarity between the feature vectors of a document, As a result of memorizing the classified result, it has the storage section and the dictionary for feature-vector generation in which the word used for a feature-vector generate time is registered. The feature vector expressing the description of each word is automatically generated using the coincidence relation between the words in a lot of text files, and it is characterized by the ability to classify a document automatically.
[0006] Moreover, in addition to the above-mentioned configuration, invention concerning claim 2 is further equipped with the useful word election section which elects a useful word using the classification result memorized by the result storage section at the time of a classification. It is characterized by the ability to raise the precision of a classification by electing a useful word as a classification and using only a useful word for it at a classification by investigating the incidence of a word for each [which was classified] of that taxon of every, after classifying a lot of text files.
[0007] Moreover, the representation vector generation section which asks for the feature vector which represents each taxon using the classification result invention concerning claim 3 is remembered to be by the result storage section in addition to the above-mentioned configuration, It has further the representation vector storage section which memorizes the representation vector generated in the representation vector generation section. After classifying a lot of text files, it is characterized by the ability to ask for the feature vector representing the field using the word for every taxon and feature vector of a document which were classified.

[Translation done.]

OPERATION

[Function] The operation at the time of study of the feature vector of the word in claim 1 is explained. The contents of a lot of text files memorized by the document storage section are passed to the document analysis section, analyses (morphological analysis etc.) of a sentence are performed, coincidence relation, the frequency of occurrence, etc. of a word are analyzed in the word vector generation section, and the feature vector of each word is generated. In this way, the feature vector of the generated word is memorized by the word vector storage section. Thus, study of the feature vector of a word is performed. The storage space of a feature vector prevents becoming huge too much with restricting the word which generates a feature vector to the word registered into the dictionary for feature-vector generation.
[0009] The operation at the time of a classification of the document in claim 1 is explained. The contents of the text file memorized by the document storage section when classifying a text are passed to the document analysis section, analyses (morphological analysis etc.) of a sentence are performed, in the document vector generation section, it asks for the feature vector of the word which appears when analyzing a sentence in the document analysis section with reference to the word vector storage section, and the feature vector of a document is generated from the feature vector of the word contained in a document. In this way, the feature vector of the generated document is memorized by the document vector storage section, and classifies a document according to the similarity between the feature vectors of this document in the classification section. This classification result is memorized by the result storage section.
[0010] With a configuration according to claim 2, after performing a classification of a lot of documents, a useful word is elected in the useful word election section using the classification result memorized by the result storage section at the time of a classification. By classifying again using the feature vector of the word which was made to learn the feature vector of a word again after registering into the dictionary for feature-vector generation only the word elected by the useful word election section, then was obtained, the storage spaces of a feature vector can be reduced rather than the configuration of claim 1, and the precision of a classification can also be raised.
[0011] With a configuration according to claim 3, it asks for the feature vector which represents each taxon with the representation vector generation section using the classification result memorized by the result storage section, after performing a classification of a lot of documents. The representation vector generated in the representation vector generation section is memorized by the representation vector storage section. Once it generates the representation vector of each taxon, when classifying new document data, it can judge to which taxon the document belongs only by performing the comparison with the feature vector of the document, and the representation vector of each taxon.

[Translation done.]

EXAMPLE

[Example] Hereafter, the suitable example of this invention is explained in full detail based on a drawing.
[0013] One example of invention concerning claim 1 is shown in drawing 1 . The storage section 101 document classification equipment remembers document data to be, and the document analysis section 102 which analyzes document data, The word vector generation section 103 which generates automatically the feature vector which expresses the description of each word using the coincidence relation between the words in a document, The document vector generation section 105 which generates the feature vector of a document from the feature vector of the word vector storage section 104 which memorizes the feature vector, and the word contained in the document, The document vector storage section 106 which memorizes the feature vector, and the classification section 107 which classifies a document using the similarity between the feature vectors of a document, As a result of memorizing the classified result, it consists of the storage section 108 and a dictionary 109 for feature-vector generation in which the word used for a feature-vector generate time is registered.
[0014] It is more realistic to restrict the word used in case a feature vector is created, since there are very many words currently generally used for the usual document. For this reason, it is the dictionary 109 for feature-vector generation, and creating the feature vector of a word only using the word registered here uses, and it can suppress growing gigantic of the storage space of a feature vector.
[0015] Drawing 2 shows the system configuration at the time of study of the feature vector of a word, at the time of study of the feature vector of a word, a lot of document data document storage section 101 for study is made to memorize, the document data read from the document storage section 101 are read into the document analysis section 102 for every suitable lumps, such as a report, a paragraph, and one etc. sentence, the document data is analyzed in the document analysis section 102, and a word is extracted. The feature vector of the word which generated the feature vector of a word in the word vector generation section 103 based on the word train extracted here, and was generated in the word vector generation section 103 is memorized by the word vector storage section 104. In this way, the feature vector of a word is learned.
[0016] When drawing 3 shows the system configuration at the time of a document classification and a document is classified, the document data which the document storage section 101 was made to memorize the data of the document to classify, and read them from the document storage section 101 are read into every [to make it classify into] unit (for example, report unit) by the document analysis section 102, the document data is analyzed in the document analysis section 102, and a word is extracted. It asks for the feature vector of the word extracted here with reference to the contents of the word vector storage section of 104. Usually, although two or more words are extracted from one unit (for example, one report) of document data, at this time, the feature vector of a document is calculated by equalizing the value of the feature vector of all the words called for. a better value may be acquired, if it does not equalize simply, but each feature vector is equalized at this time after carrying out weighting according to the inverse number of that frequency of occurrence, investigating the number of reports in which that word

has appeared from a lot of reports and hanging log (the number of reports in which the total number of reports / its word has appeared) on the feature vector of that word for example, --
[0017] If the feature vector of a document can be found, a document can be classified according to applying the technique of the conventional clustering. What is necessary is just to consider that documents with a distance near [this] between the feature vectors of a document belong to the same field.
[0018] Moreover, human being chooses a typical document for every taxon, the temporary representation vector of the taxon is generated from the feature vector of the word extracted from the document, and the feature vector of the document read from the document storage section 101 can also classify a document according to whether to be close to the temporary representation vector of which taxon. If document data are made to read from the document storage section 101 in large quantities also by such classification technique, the effect of the error that human being has chosen the temporary representation vector decreases, and, finally the quite general representation vector for each field can be generated.
[0019] Then, the method of generating the feature vector of a word is explained concretely. The feature vector of a word is obtained by adding what applied the frequency of occurrence in the inside of the document data of the lump of the word to frequency-of-occurrence distribution of the word contained in the document data of a lump. A concrete example explains.
[0020] Example A "the American government has proposed radical reexamination of the COCOM regulation to the advanced major power."
Example B "to which it seems that the country for regulation is inclined to reduce COCOM's regulated items sharply on condition that export of the industrial product which leads to manufacture of arms is regulated"
It explains how the feature vector of a word is created from the document data to say. Here although [document data] read in the unit of "one sentence", other units, such as one report, are sufficient as this.
[0021] moreover, the number of dimension of a feature vector -- 21 dimensions (the number of words registered into the dictionary for feature-vector generation is 21) -- each element -- "United States, the government, advanced, main, a country, COCOM, and regulation -- it improves and suppose that radical and the word a proposal, an object, arms, manufacture, industry, a product, export, conditions, items, large, reduction, and intention" are supported.
[0022] under such conditions, if Example A is read from the document storage section 101, the document analysis section 102 will analyze -- having -- "United States, the government, advanced, main, a country, and COCOM -- radical [regulation and] -- it improves and proposal" is extracted. At this time, 1 is added to the element corresponding to these words of the feature vectors of all these words in the word vector generation section 103. Then, the "United States", the "government", etc. add (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) to the feature vector of the word which appears in Example A. The thing illustrating this is drawing 8 .
[0023] Next, if Example B is read from the document storage section 101, it will be analyzed in the document analysis section 102, and "regulation, an object, a country, arms, manufacture, industry, a product, export, regulation, conditions, COCOM, regulation, items, large, reduction, and an intention" will be extracted.
[0024] The word frequency-of-occurrence distribution acquired from now on is (0, 0, 0, 0, 1, 1, 3, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1). It adds to the feature vector of "regulation". the vector which doubled this word frequency-of-occurrence distribution three since "regulation" had appeared 3 times -- it is (0, 0, 0, 0, 3, 3, 9, 0, 0, 0, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3) -- An "object", a "country", etc. add (0, 0, 0, 0, 1, 1, 3, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1) to the feature vector of the word which appears in Example B only once. The thing illustrating this is drawing 9 .
[0025] In addition, since the magnitude of a vector added with the die length of a sentence by the approach of always adding an integer in this way changes, how to add, after normalizing the absolute value of a vector to add to 1, or normalizing the absolute value of a vector of frequency-of-occurrence distribution to 1 and applying the value proportional to the number of appearances etc. is considered.

[0026] And the feature vector finally obtained normalizes the absolute value to 1.
[0027] In this way, the feature vector of the obtained word is memorized by the word vector storage section 104, and is used at the time of a classification of a document.
[0028] Next, the time of the following examples C being read considering processing of feature-vector generation of the document at the time of a document classification as an example is raised and explained.
[0029] Example C "the American government proposed reduction of arms."
If Example C is read from the document storage section 101, it will be analyzed in the document analysis section 102, and "the United States, the government, arms, reduction, and a proposal" will be extracted. Then, in the document vector generation section 105, with reference to the contents of the word vector storage section 104, the "United States", the "government", etc. add the feature vector of the word which appears in Example C, and get (3, 3, 3, 3, 5, 5, 9, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2) as a feature vector of Example C. The thing illustrating this is drawing 10 . In drawing 10 , since priority is given to intelligibility, the normalization of a vector is not performed, but in actual processing, after normalizing the absolute value of the feature vector of each word to 1 before adding, it adds. The obtained feature vector is memorized by the document vector storage group 106.
[0030] Next, it explains how the feature vector of a document is used in the classification section 107 at the time of a classification. After normalizing simply the absolute value of the feature vector of the document which was able to be found first to 1 Although what is necessary is to classify using a certain technique from the former, such as the K-means method, or just to classify according to similarity (obtained by finding distance or calculating an inner product) with the representation (temporary) vector of a taxon Since the feature vector obtained by this technique has the description "the value of the element corresponding to the word appearing [many] becomes very large", a classification result with it better [to devise so that this description may not have a bad influence on a classification] is obtained in many cases. For example, distance by count which the difference between elements does not expand in finding distance (although the square root of the sum of squares of the difference between each element is usually calculated) For example, it is good to normalize, after taking log for all elements or taking a power root, before it is better to have used the distance which calculated and asked for the sum of the absolute value of the difference between each element and asking for an inner product, and to calculate, after leveling a value.
[0031] The juniper currently asked for those with three, and the representation vector of each taxon for the taxon as follows as an example of a classification.
[0032] The representation vector of a taxon 1 (1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1)
The representation vector of a taxon 2 (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5)
The representation vector of a taxon 3 (4, 4, 4, 4, 6, 6, 6, 3, 3, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
After the feature vector of a document and the representation vector of a taxon normalize an absolute value to 1, supposing what calculates both inner product and takes the biggest value is the highest as a scale of similarity as for similarity, it is the feature vector [0033] of Example C.
[Equation 1]
$$\frac{1}{\sqrt{238}}$$

[0034] (3,3,3,3,5,5,9,3,3,3,2,2,2,2,2,2,2,2,2,2,2)
The representation vector of a taxon 1 [0035]
[Equation 2]
$$\frac{1}{\sqrt{8}}$$

[0036] (1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1)
The representation vector of a taxon 2 [0037]
[Equation 3]
$$\frac{1}{\sqrt{285}}$$

[0038] (1,1,1,1,1,1,1,1,1,1,5,5,5,5,5,5,5,5,5,5)
The representation vector of a taxon 3 [0039]
[Equation 4]
$$\frac{1}{\sqrt{210}}$$

[0040] (4,4,4,4,6,6,6,3,3,3,1,1,1,1,1,1,1,1,1,1)
since -- the inner product of the feature vector of Example C, and the representation vector of each taxon -- an inner product [0041] with a taxon 1
[Equation 5]
$$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{8}} \cdot 20 \simeq 0.4583$$

[0042] An inner product with a taxon 2 [0043]
[Equation 6]
$$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{285}} \cdot 150 \simeq 0.5759$$

[0044] An inner product with a taxon 3 [0045]
[Equation 7]
$$\frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{210}} \cdot 211 \simeq 0.9438$$

[0046] Since it turns out that the feature vector of a next door and Example C is the closest to the representation vector of a taxon 3, Example C is classified into a taxon 3. Drawing 11 illustrated this. Since priority is given to intelligibility like [drawing 11] drawing 10 , the normalization of a vector is not performed, but in actual processing, after normalizing each absolute value of a vector to 1 before comparing, it compares. The classified result is memorized by the result storage section 108.
[0047] Next, one example of claim 2 of this invention is shown in drawing 4 . Here, what is expressed with signs 201-209 is the same as what is expressed with the signs 101-109 of drawing 1 respectively.
[0048] The storage section 201 document classification equipment remembers document data to be, and the document analysis section 202 which analyzes document data, The word vector generation section 203 which generates automatically the feature vector which expresses the description of each word using the coincidence relation between the words in a document, The document vector generation section 205 which generates the feature vector of a document from the feature vector of the word vector storage section 204 which memorizes the feature vector, and the word contained in the document, The document vector storage section 206 which memorizes the feature vector, and the classification section 207 which classifies a document using the similarity between the feature vectors of a document, As a result of memorizing the classified result, it consists of the storage section 208, a dictionary 209 for feature-vector generation in which the word used for a feature-vector generate time is registered, and the useful word election section 210 which elects a useful word using the classification result memorized by the storage section 208 the result at the time of a classification.
[0049] Drawing 5 is drawing showing the system configuration at the time of study and a classification. At first, the feature vector of a word is learned and a lot of document data are classified according to the same approach as the example of claim 1 based on it. Although the classified result is memorized by the result storage section 208, it carries out based on this result, and a useful word is elected in the useful word election section 210. This asks for the frequency of each word for every taxon, removes the word contained at the same rate as every taxon, or elects a ********** thing for the following [a threshold with the ratio of the (approach 1:highest frequency and the minimum frequency] only as removal) and a certain taxon at a high rate (approach 2: elect the thing beyond the highest frequency and a threshold with the second place of a ratio with frequency). In addition, the word which elects in the useful word election

section 210 may not necessarily be from the word registered into the dictionary 209 for feature-vector generation, and can perform election from the word of the larger range.
[0050] As an example, a taxon presupposes that the word which are a, b, and c and which is registered into the dictionary 209 for feature-vector generation was three "politics, Japan, and international" noting that there are three. And the frequency of each word (suppose that frequency is investigated also about "Election" and a "problem" in addition to the word registered into the dictionary 209 for feature-vector generation) presupposes that it was as follows for every taxon.
[0051] Taxon a 30% of politics, Japan 5%, 35% of international, 10% of Election, 20% taxon b of problems Politics 3%, 55% of Japan, 35% of international, Election 2%, problem 5% taxon c Politics 3%, 30% of Japan, 35% of international, Election If it carries out 30% of problems 2% Since "international" is contained at the same rate as every taxon if an approach 1 is used, it will remove from the dictionary for feature-vector generation. Since "politics", "Japan", "Election", and a "problem" have a bias in the frequency for every taxon, it is elected as a useful word and registers with the dictionary 209 for feature-vector generation (what is necessary is to take only the number to register in order of the total frequency of occurrence in the word which has a bias in frequency to stop the number of registered words at this time). When an approach 2 is used, "politics" and "Election" are elected and it registers with the dictionary 209 for feature-vector generation, and "Japan", international ["international"], and a "problem" are not registered into the dictionary 209 for feature-vector generation. The method of electing a useful word by whether it is beyond a threshold with the ratio of the frequency of the 1st place and the frequency of the n-th place (n is the number of 3 or more and a taxon - one or less) as the in-between approach of an approach 1 and an approach 2 is also considered. Moreover, the method of electing what has large not the ratio of frequency but value of distribution of frequency is also considered.
[0052] In addition, since it is possible that the word elected by doing in this way has the significance according to the ratio (or distribution of frequency) of frequency, if the feature vector of the word in that document is equalized after carrying out weighting according to this ratio (or distribution) when calculating the feature vector of a document (after hanging log (ratio of frequency) on that feature vector), the feature-vector ground of a better document may be obtained.
[0053] In this way, only a useful word is registered into a classification, once again, if the feature vector of a word is learned and a document is classified using it, the dictionary for feature-vector generation can be made smaller, or the precision of a classification can be raised to the dictionary 209 for feature-vector generation.
[0054] One example of claim 3 of this invention is shown in drawing 6 . Here, what is expressed with signs 301-310 is the same as what is expressed with 201-210 of drawing 4 respectively.
[0055] The storage section 301 document classification equipment remembers document data to be, and the document analysis section 302 which analyzes document data, The word vector generation section 303 which generates automatically the feature vector which expresses the description of each word using the coincidence relation between the words in a document, The document vector generation section 305 which generates the feature vector of a document from the feature vector of the word vector storage section 304 which memorizes the feature vector, and the word contained in the document, The document vector storage section 306 which memorizes the feature vector, and the classification section 307 which classifies a document using the similarity between the feature vectors of a document, The dictionary 309 for feature-vector generation in which the storage section 308 and the word used for a feature-vector generate time are registered as a result of memorizing the classified result, The useful word election section 310 which elects a useful word using the classification result memorized by the result storage section 308 at the time of a classification, It consists of the representation vector generation section 311 which asks for the feature vector which represents each taxon using the classification result memorized by the result storage section 308, and the representation vector storage section 312 which memorizes the representation vector generated in the representation vector generation section 311.

[0056] In addition, in constituting the system of claim 3 using the example of claim 1, it becomes a system without the useful word election section 310.
[0057] Drawing 7 is drawing showing the system configuration at the time of study and a classification. At first, the feature vector of a word is learned and a lot of document data are classified according to the same approach as the example of claim 1, or the example of claim 2 based on it. Although the result storage section 308 memorizes, the classified result is carried out based on this result, and generates a representation vector in the representation vector generation section 311. this asking for the frequency of each word for every taxon, electing the word contained only in a certain taxon at a high rate, and taking the average of the feature vector of such a word -- it is generable. As an example, a taxon presupposes that the word which are a, b, and c and which is registered into the dictionary 309 for feature-vector generation was three "politics, Parliament, and international" noting that there are three. And the frequency of each word for every taxon presupposes that it was as follows.
[0058]
Taxon a 40% of politics, 50% of Parliaments, 10% taxon b of international 10% of politics, 10% of Parliaments, 80% taxon c of international If it carries out 20% of politics, 10% of Parliaments, and 70% of international, the representation vector of Taxon a will be given as an average of a "political" feature vector and the feature vector of "Parliament." In addition, giving weight is also considered by not a mere average but the appearance rate. For example, if the "political" frequency of occurrence is twice the frequency of occurrence of "Parliament", it is making into the representation vector of Taxon a what added the twice of a "political" feature vector, and the feature vector of "Parliament", and was divided by 3 etc.
[0059] How to make what took the average of the feature vector of the document similarly classified into Taxon a the representation vector of Taxon a is also considered.
[0060] In this way, the document read from the document storage section 301 can be classified now into the taxon corresponding to a representation vector most similar to the feature vector of that document according to referring to this representation vector at the time of a classification of future documents by memorizing it in the representation vector storage section 312, if a representation vector is generated.
[0061] This invention classifies an electronic mail and electronic news automatically, or it not only uses it for a document classification, but Elect what is likely to have a user's interest out of an electronic mail and electronic news, or (A user can judge by similarity with the feature vector of the mail read by then or news) Ambiguous retrieval (by searching the document which becomes beyond a threshold with the fixed similarity of the feature vector of a retrieval keyword, and the feature vector of the document for retrieval) Even if it does not match a retrieval keyword correctly, can use for the ability to refer to the keyword of relation, or Can use for selection (the homophenes is chosen by similarity with the feature vector obtained from the contents changed by then) of the homophenes in the conversion of kana into kanji, or Also in case the approach of choosing the conversion result of having suited the past context most in speech recognition, handwriting recognition, etc. is taken (a recognition result is chosen by similarity with the feature vector obtained from the contents recognized by then), can use, or It can use, also in case retrieval space, such as a word, is narrowed in the time of recognition etc. (only the word corresponding to the element which has become among the elements of the feature vector obtained from the contents recognized by then beyond the fixed threshold is searched).

[Translation done.]

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is the block diagram showing the basic configuration of one example of invention concerning claim 1.
[Drawing 2] It is the block diagram showing the system configuration at the time of study of the system shown in drawing 1 .
[Drawing 3] It is the block diagram showing the system configuration at the time of a classification of the system shown in drawing 1 .
[Drawing 4] It is the block diagram showing the basic configuration of one example of invention concerning claim 2.
[Drawing 5] It is the block diagram showing study of the system shown in drawing 4 , and the system configuration at the time of a classification.
[Drawing 6] It is the block diagram showing the basic configuration of one example of invention concerning claim 3.
[Drawing 7] It is the block diagram showing study of the system shown in drawing 6 , and the system configuration at the time of a classification.
[Drawing 8] It is drawing explaining generation of the feature vector of a word.
[Drawing 9] It is drawing explaining generation of the feature vector of a word.
[Drawing 10] It is drawing explaining generation of the feature vector of a document.
[Drawing 11] It is drawing explaining a classification of a document.
[Description of Notations]
101, 201, 301 Document storage section
102, 202, 302 Document analysis section
103, 203, 303 Word vector generation section
104, 204, 304 Word vector storage section
105, 205, 305 Document vector generation section
106, 206, 306 Document vector storage section
107, 207, 308 Classification section
108, 208, 308 Result storage section
109, 209, 309 Dictionary for feature-vector generation
210 310 Useful word election section
311 Representation Vector Generation Section
312 Representation Vector Storage Section

[Translation done.]